DOCUMENT RESUME

ED 069 706 TM 002 158

AUTHOR	Gulliksen, Harold
TITLE	Looking Back and Looking Ahead in Psychometrics.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RM-72-8
PUB DATE	Jul 72
NOTE	30p.; Paper presented at the spring meeting of the Psychometric Society (Princeton, N.J., March, 1972)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Bayesian Statistics; *Bibliographies; Factor Analysis; *Historical Reviews; Learning Theories; *Measurement Techniques; *Psychometrics; Rating Scales; Speeches
IDENTIFIERS	Differential Aptitude Tests

ABSTRACT

A presentation of the 40-year history of psychometrics is given with comments about needed trends for the future. Computers have radically changed the time required for data processing. In testing, many promising developments, such as Kristof's reliability for vector variables, latent class and latent struction models, one-factor ration scale in testing and Bayesian procedures, are still largely in the theoretical field. Interest in scaling did not become important until Messick applied methods previously developed to attitude scales in 1956. Multidimensional scaling techniques have recently been utilized in a number of research areas and applied fields. Factor analysis theories are reasonably well developed. Applications to aptitude tests have been made, but have been only sketchily used in other fields in which they would be extremely valuable, such as economics, sociology, and physiology. In the field of mathematical learning theory, work needs to be done for individual learning curves and in comparing various stochastic and continuous models. Quantative psychology has moved a long way in 40 years. (DJ)

ED 069706

RM-72-8

RESEARCH MEMORANDUM

LOOKING BACK AND LOOKING AHEAD IN PSYCHOMETRICS

Harold Gulliksen

TM 002 158

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

This is a revised version of an after dinner talk given on March 30, 1972, at the spring meeting of the Psychometric Society, Princeton, New Jersey.

Educational Testing Service
Princeton, New Jersey
July 1972

Looking Back and Looking Ahead in Psychometrics[1]

Harold Gulliksen

In presenting the history, I shall do so in terms of the various investigators who worked in this area, even though I would, in general, agree with James Thurber in doubting the great man theory of the development of science. He points out that some people think it was a great day, and a critical event when Benjamin Franklin sent up his kite and brought down lightning, demonstrating its fundamental similarity to electricity (Thurber, 1937). Others feel, however, that this event was not particularly critical, believing that if Franklin had not done this, somebody else would have made the same discovery; and we can see that this was exactly what happened with the harnessing of steam and the invention of the gas engine. Franklin didn't make these discoveries, and sure enough somebody else did. Q.E.D.

"Looking back in Psychometrics" for me goes to 1929 when I was a graduate student at Ohio State University and Thurstone gave a seminar on the theory and applications of the law of comparative judgment and the method of paired comparisons. Numerous topics, such as art, esthetics, ethics, subjective values, etc., had previously been dismissed with "What can you do about field or topic X? It is all a matter of opinion, and opinions disagree." I was tremendously impressed by the idea that now there was a clear-cut theory and experimental procedure for a rigorous treatment of those areas that are entirely a matter of opinion, and it was essential for the use of the method that the opinions should disagree (see Thurstone, 1959, for a collection of his articles on scaling).

---

[1]This is a revised version of an after dinner talk given on March 30, 1972, at the spring meeting of the Psychometric Society, Princeton, N. J.

Following the advice of Albert Paul Weiss, the senior professor in experimental psychology at Ohio State, I attended the University of Chicago summer school in 1929, and took a six-week course with Thurstone in which he covered test theory, scaling, factor analysis, and I believe, mathematical learning theory. A week or two was spent on each of the four topics, and that was that.

A decade earlier E. L. Thorndike perceived the necessity and possibility for such developments in quantitative psychology--"Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality" (Thorndike, 1918, page 194).

Computers

Indicating the developments of the last 40 years requires mention of the electronic computer. I was a research assistant for a year working on Thurstone's first study of primary mental abilities. The computational work in resolving a battery of about 50 tests into seven primary mental abilities meant that I was supervising a group of about 20 computer clerical workers for about a year. I recall Thurstone lamenting that his Ph.D. candidates would not be able to do factor analysis dissertations because it would not be practical to employ such a crew for each Ph.D. thesis. A few years ago, a research worker in the Civil Service in Washington, D. C., wanted some help in analyzing a set of attitude scales which he had given to different types of persons working under Civil Service, in order to see how the jobs could be changed to make them more attractive. He came up one afternoon, with his data on punched cards, and we started about 4:00 in the afternoon to run the preliminary error detecting program and we corrected the cards whenever errors

were found. In all, including the scaling, correlations, factor analyses, and rotations, although the job was somewhat larger than the primary mental abilities one, we were finished about 3:00 the next morning.

Testing

During the past 40 odd years we have come a long way, as you all know, since the publication of Thurstone's (1931c) first test theory text. In reliability theory the widely used K-R 20 and K-R 21 were developed (Kuder & Richardson, 1937), and we have now progressed to Kristof's (1972) reliability for vector variables.

Latent class and latent structure models have been developed by Birnbaum, Lazarsfeld, and Bert Green among others. Rasch has presented a theory for a one-factor ratio scale in testing. Mel Novick, Charles Lewis, and others have worked on Bayesian procedures. These and other developments have been presented and summarized by Lord and Novick (1968) in Statistical Theories of Mental Test Scores. The theory and practical applications of tailored testing are being investigated by Lord (1971), Cronbach and Gleser (1965), and others.

The foregoing developments, however, are still largely in the theoretical field. I hope they will have greater impact on the standard aptitude and achievement tests. As far as I am aware, there has been little or no impact on teacher constructed tests used in grading classes. Instead, there is today what seems to be a serious movement away from any type of measurement in education, rather than an attempt to use better measurement methods.

In aptitude test development, the rule is to take validity coefficients seriously without raising the question: "Should this validity be high or

low?" During World War II, while working on aptitude and achievement test development for the Navy, Norman Frederiksen and I obtained considerable experience in this area. At the Gunners' Mates school, we found that the validity of the reading test was high, and the mechanical knowledge and mechanical comprehension tests had low validity. We worked for about six months and developed identification and performance tests that measured the objectives given to us by the Gunners' Mates school. On the basis of grades on the new achievement testing program, the validity of the reading test took a nose dive, and the mechanical comprehension and mechanical knowledge went up. The same thing happened in basic engineering, where the arithmetic test showed highest validity initially. Nicholas Fattu worked for a year developing gauges to measure the products quickly and accurately, and, on the basis of the achievement measures, the validity of the arithmetic test dropped and that of the mechanical aptitude tests went up. Similar results were obtained in the Torpedoman's and other schools. Some of this work has been written up in Stuit (1947), especially chapters XII, XIII, and XV.

I think school and college grades are in need of similar scrutiny. For example, the spatial relations test of the College Board showed good validity for grades in some engineering drawing classes and poor validity for other engineering drawing grades. Such results would be expected if these courses, which were all given the same name--engineering drawing--were in fact quite different, and were graded on different bases. A general discussion of such problems under the title of "Intrinsic Validity" can be found in Gulliksen (1950).

A paper by Plotkin in the March 1972 issue of the American Psychologist discusses problems in the area of the validity of tests brought to the fore by the Equal Employment Opportunities Act of 1964.

I feel sure that Ben Shimberg and his associates, who are working at ETS on such things as tests for auto mechanics, will not accept the conclusion that the major important quality for an automobile repairman is high verbal ability, on the basis of validity studies, but will see to it that the criteria are changed so that the important abilities are mechanical skill, trouble shooting ability, etc.

Some years ago, in looking over validity coefficients for the Differential Aptitude Test, I noticed that for one school the best predictor of grades in Latin was the clerical test (.47). For the other tests of the Differential Aptitude Test, the correlations with Latin grades ranged from a low of -.37 for mechanical reasoning through -.02 for verbal reasoning, to a high of .19 for sentences. It was pleasing to note that this was not generally true for all the schools studied. But it would be even more pleasing, if some steps had been taken to alter the teaching and grading procedures in that school. Other studies by the Psychological Corporation showed that higher educational level goes with higher clerical ability (see Bennett, Seashore, & Wesman, 1959, pp. 48, 79; 1966, pp. 5-42).

In 1939, Truman Kelley wrote on "Mental Factors of No Importance," noting that only the verbal and quantitative abilities seem to be important as far as academic work in schools and universities is concerned. He spoke of the numerous abilities which even then were being isolated by factor analysis and indicated his fear that "many of the factors thus far 'found' approach

pretty close to the limit of no importance." This may well be true of many of the 150 or so factors in the French (1951) monograph, but before reaching any such conclusions, the school's teaching, testing, and grading procedures should be studied carefully and revised where necessary. My own judgment would be that when this is done properly we will find that verbal and quantitative do not exhaust the list of useful abilities.

In 1901, Clark Wissler, while getting his Ph.D. with James McKeen Cattell at Columbia, investigated the validity of a number of tests for predicting grades at Columbia. The validities ranged from -.02 for reaction time to .19 for logical memory. During the seven decades since then, aptitude and standardized achievement tests have advanced tremendously. However, I think the evidence is that college grades are now about the same as they were at the turn of the century. Wissler (1901) reported that the correlations of grades ranged from a low of .30 for Rhetoric and French to a high of .75 for Latin and Greek, which I believe would be very similar to the correlations obtained today.

During recent decades there have been a few attempts at improving the quality of college exams and grades. An example is the work of the examining office at the University of Chicago during the 1930's and 1940's under the direction of L. L. Thurstone and Ralph Tyler. The Chicago faculty later abandoned this program.

The great need we have now is not for the improvement of aptitude tests, but for improvement in the criteria against which they are evaluated, including not only grades for four years in college, but activities and achievements during the 40 years after college.

Scaling

In the early 1930's, Thurstone's interests changed from scaling to factor analysis. For the next decade development in the scaling area was very slow. Marion Richardson and I asked Gale Young and Alston Householder about the problem of determining dimensionality and a coordinate system for a set of points from the interpoint distances. They solved the problem and published the Young and Householder paper on multidimensional scaling, "A Discussion of a Set of Points in Terms of Their Mutual Distances," in 1938. The applications of this method by Marion Richardson and Klingberg appeared at about the same time. Otherwise, not much appeared until Torgerson developed the theory and verified a section of the Munsell color system in his thesis in 1951. Messick (1956) applied the method to attitude scales a little later. Since then the development of scaling theory has been tremendous: the law of categorical judgment, the method of successive intervals, and statistical tests for fit of data to theory. These developments are presented in Torgerson's (1958) Theory and Methods of Scaling. Since then there has been the extension to take care of individual differences, which makes the methods far more useful in attitude measurement and other applications in social psychology (see Carroll & Chang, 1970; Helm & Tucker, 1962; Tucker, 1972; Tucker & Messick, 1963). Luce and Tukey (1964) have developed the theory of conjoint measurement and shown the independent foundation on which psychological measurement rests. Bock and Jones (1968) have given rigorous estimation procedures. The theory has been developed by Suppes and Zinnes, Tversky and others, so that measurements from psychological scaling are not dependent on other methods (see Luce, Bush, & Galanter, 1963, 1965). Indow and his group in Japan, Ekman

and his group in Sweden, and Stevens in the United States have been active in the development of theory and applications of scaling.

Applications of scaling techniques to linguistics and free recall are illustrated by the work of John B. Carroll (1971) and Friendly (1972). My own work (Gulliksen & Gulliksen, 1971) has also illustrated the application of scaling and factor techniques to attitudes toward work and leisure in cross-cultural comparisons. Coombs (1964) and his co-workers have developed a nonmetric multidimensional unfolding procedure, using this method to study confusions of Morse code signals, and showed two dimensions in the subset of 10 signals studied.

While listening to the Psychometric Society papers here today, I was strongly reminded of Stephen Leacock's (1911) energetic young lord who flung himself on his horse and rode off in all directions at once.

As indicated above there have been numerous applications of scaling techniques by research workers in various academic university settings. However, when we consider the various applied fields in which linear and multidimensional scaling could be used, the picture is different from the development of theory and applications in the academic setting. The various polling organizations report nothing but total percentages, sometimes broken down by various preformed categories, such as education, sex, rural-urban, etc. I have never seen a single instance where a factor analysis of a set of observations is given, so that various points of view, or clusters of opinions, can be found. Bob Tryon (1955) reported a factor analysis of voting areas around San Francisco, and found that the various indices available formed a three-factor system. He suggested that voting might well be associated with these factors, so that one would get better prediction by repeating such a study in connection with a new election poll and using these factors as independent variables in

adjusting the polling results. As far as I am aware, no polling group has paid any attention to such possibilities.

Green and Carmone (1970) and Green and Rao (1972) have shown how multidimensional procedures could give valuable information in consumer surveys. Applications in behavioral sciences have been given in Shepard, Romney and Nerlove (1972). Applications in marketing research have been presented by Bass, King and Pessemier (1968). The use of scaling methods in studies in perception by Carroll and others are reported in Carterette and Friedman (1973). Bell laboratories has compiled a bibliography of recent studies and applications of multidimensional scaling (Harris, 1972). It is pleasing to note that the scaling techniques have recently been utilized in a number of research areas and applied fields.

Factor Analysis

With respect to factor analysis, the initial papers presenting the principal components and other methods were published by Thurstone (1931a, b; 1933) and Hotelling (1933), following earlier work by Spearman and Holzinger. Thurstone presented his problem to Bliss in mathematics and Bartky in statistics one noon at the Chicago Quadrangle Club. He explained that he had a square symmetric array of numbers and wanted to express it in terms of summed products of a smaller array. Their reaction was, "Oh, you mean the square root of a symmetric matrix." In this way Thurstone learned that matrix theory existed and was relevant to the factor problem, so he embarked on a year or two of tutoring and published The Vectors of Mind (1935), followed later by Multiple Factor Analysis (1947), giving a concise summary of the crucial aspects of matrix theory and their use in factor analysis. Prior to the advent of

electronic computers, approximations such as the centroid method, with largest correlations used as initial estimates of communalities, were widely used because of their practicality.

I remember Sam Wilks remonstrating about this. He said, "We know a good method, the principal components, based on least squares, that gives a best fitting reduced rank matrix. Why can't you use that instead of these ad hoc approximations whose properties are unknown?"

Since then Lawley, and Jöreskog (1970) with Gruvaeus and van Thillo have presented the theory and associated practical computer methods. Kaiser's (1970) little jiffy is very widely used. Tucker has given us the procedures for double centered matrices (Tucker, 1956), for the inter-battery matrix (Tucker, 1958) and for three and multi-mode analysis (Tucker, 1966a). Harris (1962, 1963) has presented relations among factor theories and cautions that should be observed when attempting to measure change. Guttman (1971) has presented some extensions and applications of his facet theory. Horst (1961) has given possible applications of generalized canonical correlations. McDonald (1963, 1967) has given us his nonlinear factor analysis. Arbuckle (1970) has developed a procedure using the Toeplitz matrix as the error matrix instead of a diagonal matrix, so that the factor procedures may be applied to matrices where it is reasonable to assume "stationary error," as in analyzing nerve potentials.

Factor analysis theory, and associated electronic computer programs, are in a reasonably well developed state. As to applications of the methods, it has been mentioned previously that numerous aptitude test batteries have been analyzed, so that psychologists have some reasonable notions regarding the basic abilities represented in aptitude tests.

The factor methods have, however, been only sketchily used in other fields where they would be extremely valuable. Harman (1967) devotes about a page of his text to indicating applications in economics, sociology, physiology, etc., but the impact of these factor studies on the fields indicated has been minimal.

Schiffman and Falkenberg (1968) have presented an interesting study of matrices with stimuli designating rows, by retinal cells designating columns; or stimuli by taste neurones, that give interesting pictures of the structure of these sensory systems. In the study of retinal cells, the stimuli spread out in a curve--violet, blue, green, yellow, orange, and red--while in the same space the retinal cells clustered three in the blue area, four in the green, and four in the red. For taste a definite three-dimensional structure was obtained, but the details are not so clearly interpretable.

Memory is another field in which factor analysis would be extremely valuable. Paul Kelley's (1964) study, for example, demonstrated that memory span is a factor that includes visual and auditory material, as well as nonsense and meaningful material. However, when one deals with longer lists, so that it takes a number of repetitions to memorize them, then rote memory differentiates from memory for meaningful material. That is to say, when Ebbinghaus introduced the nonsense syllable, as he thought to simply control for the irrelevant factor of possible differences in previous associations, he was unwittingly shifting to measurement of a different ability. Recently there has been great emphasis on what is termed "free recall," which means that the material, though meaningfully organized, is presented randomly, to see the extent to which the subjects will make use of the organization in recall. Again, as with Ebbinghaus, it is assumed that this is simply another

interesting procedure for tapping the memory function. However, as far as I am aware, no factor studies have been made including both the free recall, the rote, and the meaningful memory where the order of presentation must be the order of recall. We do not know whether the free recall ability is the same as the previously established rote or meaningful memory, or whether a new ability has been introduced with this new procedure. Stake (1961) has evidence indicating the possibility that the change from free study such as Ebbinghaus (1885) used, to the memory drum, introduced by Müller and Schumann (1894) merely as an added experimental control, may have altered the ability being measured.

Eight indices of "excitatory potential" (a useful hypothetical construct) were used by Lloyd Humphreys (1943) in a conditioned eyelid experiment. He found two factors. Acquisition amplitude and extinction amplitude loaded on one factor, while acquisition and extinction latencies loaded on a different factor, along with extinction frequency. Acquisition frequency loaded equally on both factors. That is to say, the experimenter's selection of one or another from a set of possible indices may really be changing the hypothetical construct being measured.

For decades it has been asserted that "intelligence is the ability to learn" (e.g., Binet, 1909, especially p. 146; Buckingham, 1921, especially p. 273; Dearborn, 1921; Peterson, 1926, especially pp. 268 & 276; Pyle, 1921). This view that intelligence is the ability to learn has been critically examined (see Woodrow, 1946; Peterson, 1926; Simrall, 1947, for example). Psychologists have devised numerous clever tasks in rote learning, meaningful learning, concept learning, motor learning, etc., such as mirror drawing, pursuit rotor, reversal learning, etc., which require an ability to learn,

that can be measured by time, errors or trials taken to reach some criterion, or by parameters of some learning curve fitted to the data. In only a few cases have such studies also included a few of the standard test scores that may be related to intelligence.

Studies in this area by Duncanson, Stake, Allison, Manley, Games, and Bunderson, reviewed in Bob Gagné's (1967) Learning and Individual Differences, have indicated that there are a number of different abilities represented by the different learning tasks, as well as a number of different abilities represented by the intelligence or ability tests. So far some of these abilities seem to be unique to measures of learning, or to test scores; but there are some factors that have loadings on both the learning scores and the test scores. A clear answer as to the relation between aptitudes as measured by tests, and learning abilities as measured by various learning tasks devised by psychologists, is not available at present. The topic is in need of much further research.

For the last 40 years there has been a profusion of factor analyses of batteries of aptitude tests, but there are numerous other areas in learning, memory, physiology, nerve potentials, economics, political science, sociology, etc., where factor analysis would be extremely valuable, and where only a few studies have been made.

Learning

With respect to mathematical learning theory, Thurstone (1930a) presented a mathematical derivation of an equation of the learning curve based on an urn analogy, which turned out to also be derivable from Thorndike's Law of Effect (Gulliksen, 1934). Thurstone (1930b) also showed how this theory could

be applied to determining a functional relationship between learning time and length of task, and to separating learning ability of the individual from the difficulty of the task, using factor methods. I presented an analytical procedure that separated learning ability from initial performance (Gulliksen, 1942).

During the 1950's a variety of learning models were presented based on stimulus sampling ideas (Estes), on stochastic processes (Bush & Mosteller, 1955), on stepwise increases or decreases in strength of correct and incorrect responses (Audley & Jonckheere, 1956) and various models suggested by Bower, Trabasso, Atkinson, Suppes, and others.

In general these more recent models tended to have two characteristics.

(1) Response strengths, or response probabilities, changed by finite amounts with each trial. The substitution of differentials for deltas was believed to be an extremely inappropriate step that must be avoided.

(2) In order to obtain good parameter estimates, it was usually assumed that all subjects in a group could be regarded as giving estimates of the same parameter values so that the record of the group of learners was analyzed to determine one set of parameters.

There are several questions introduced here that it seems to me should be subjected to careful experimental investigation, rather than being settled by assumption.

(1) We now have a variety of stochastic, or finite step models, and also older continuous models. Both of these types of models should be tried out on various types of learning data. It is perfectly possible that different theories will be best for different types of tasks. The same type of theory may well not fit conditioned escape response, maze learning, visual shape

discrimination--with attention to transposition, paw retraction to avoid a shock, conditioned emotional reactions such as rapid breathing, etc. For example, there is evidence that conditioned paw retraction does not transfer from the right to the left brain in split brain animals, while increased breathing rate transfers very rapidly. We need now a large number of studies in which various stochastic and continuous models are tried out on various types of learning data.

(2) I feel that the primary stress should be on using learning parameters that are psychologically meaningful. By this I mean parameters such as difficulty of task, learning ability, initial preference, and final performance on the task. These parameters seem to me to be meaningful in understanding differences between learning tasks, and differences between individuals in learning these tasks. Parameters such as number and length of runs of errors, average and variance of learning parameters, number of alternations, that have been frequently or usually used with stochastic models, seem to me to be parameters selected because they fit with the stochastic models, rather than because they have any interesting psychological significance in understanding the learning process or the differences between learning tasks and between learners.

Bush and Mosteller (Chapter 15 in Bush & Estes, 1959) have given a comparison of eight different learning models, with respect to their agreement with Solomon and Wynne's data on shock avoidance by 30 dogs given 25 trials each. The comparisons are entirely in terms of means and standard deviations of distributions of a number of variables, such as number of trials before first and second avoidance, total number of shocks, number of alternations, number of trials before the first run of four avoidances, etc. It is assumed either that the basic parameters are the same for all animals,

or else that the parameter varies according to some specified distribution. This seems to me to be an approach dictated basically by the characteristics of the stochastic approach, rather than by the psychologically interesting properties of learners and learning tasks. Determination of parameters for each learner offers a much better way of understanding the learning process, in terms of parameters of individuals, and parameters of the tasks, such as initial ability and learning ability, and difficulty of the task.

(3) Merrell (1931), Sidman (1952), and Estes (1956) pointed out the difficulties involved in using group or average learning curves, yet obtaining a single set of parameters for the average learning curve is still a very usual procedure. Is it legitimate to regard learning parameters as the same for all subjects in a group, or do some subjects have definitely better learning ability, or initial performance than others do? Again the answer may be different with different types of learning problems and with variations in difficulty of problem. There are at least two possible approaches that should be tried on this problem of individual differences in learning parameters. One approach proposed, and tried out on some sets of data by Tucker (1966b) and by Weitzman (1963), is a principal components analysis of a matrix of learning curves. The method gives a set of k learning parameters for each individual, and k generalized, or master learning curves. In the special case where k is equal to one, then it is legitimate to use the group or average learning curve.

(4) Parameter estimation for individual learning curves for the finite step models is a difficult problem. Procedures have been devised by Ramsay (1970), Wainer (1968), and Best (1966), for parameter estimation for individual

learning curves. Using Monte Carlo data with known parameters, Ramsay (1970) found that the input parameters were not recovered except for the limited case of only two parameters, initial probability of a correct response, and the effect of reward of a correct reponse on the strength of the correct response. Ramsay also felt that negative parameters, which allowed a decrease in response strength, should not be permitted because this might lead to negative response probabilities. Best's (1966) procedure allowed for the possibility that the strength of the incorrect response would be decreased by punishment for an error. If the fitting problem is satisfactorily solved, then various stochastic and continuous models could be compared with respect to parameter determination for individual rather than group learning curves.

(5) One of the great handicaps in the study of learning has been the impossibility of obtaining evidence on reliability by replication. When a learning curve has been obtained, a second one on the same problem for the same individual is impossible, because he already knows the solution, and cannot learn it again. If one tries a different problem, there is the question of how similar the two problems are, and also the question of positive or negative transfer. If one tries the same problem with another individual, then there are the possibilities of different learning abilities for the different individuals. Sperry (1961, esp. p. 1753; 1964, esp. p. 48) felt that his work with split brain preparations offered opportunity for replication from left to right brain with what was essentially a duplicate subject. So far the evidence here is conflicting. Meikle, Sechzer and Stellar (1962), working with cats suspended in a harness and learning to lift the front paw

to avoid a shock signalled by stroking the shoulder, found (for three animals) a very good linear relation between number of trials to criterion for right and left brain learning. Phil Best (1966) analyzed visual discrimination data from an experiment by Meikle and Sechzer (1960) and found a strong linear relation between first and second side learning in split brain cats. By contrast Ian Steele Russell (Russell & Kleinman, 1970), working with functionally split brain rats on a conditioned escape response, found that for a given difficulty of problem there was a zero correlation between trials to criterion for left vs. right brain. He points out that this is consistent with the view of learning as a finite step process. Recent work by me and Voneida also found marked dissimilarity between right and left brain learning in split brain cats.

Another problem is raised by the probability learning situation. When, for example, one stimulus is rewarded 70% of the time and the other rewarded 30% of the time, some investigators report that the subjects choose the stimuli about 70% and 30% respectively. This behavior is known as "matching" and would result in (.7 x .7) plus (.3 x .3) equals .58 success. Choosing the "70%" stimulus all of the time would result in 70% success. This is known as maximizing behavior. Stimulus sampling theory predicts matching. However, maximizing is frequently found. Wainer (1968) found that maximizing behavior was the rule, and succeeded in modifying the stimulus sampling theory so that with different parameter values it would predict either maximizing or matching. His data gave good agreement with the generalized theory and showed maximizing rather than matching. He devised methods of fitting parameters to individual curves.

Richard Rose (Rose, Beach, & Peterson, 1971) at the University of Washington has recently reviewed the major studies in this field and concludes that "probability matching though widely accepted by psychologists is not found when individual records are examined, instead of group averages. The individual response probabilities are much further away from matching or other theoretical values than would be permitted by the most generous interpretation of extant theories." In my view this points to the desirability of estimating parameters for individual rather than group curves.

In the field of mathematical learning theory, it seems to me that a great deal of work still needs to be done on parameter estimation for individual learning curves and in comparing various stochastic and continuous models. By contrast, in the fields of test theory, scaling, and factor theory, the theory including parameter estimation, significance testing, and variance components analysis procedures are reasonably well developed. Test theory, though adequately utilized in standardized testing programs, has not yet had much impact on the teacher constructed tests and on grading procedures. Recently scaling, especially multidimensional scaling, has received considerable attention from workers in certain applied fields and in some research areas, but its use could be more widely extended, as for example in election polls. Numerous batteries of aptitude tests have been factor analyzed--but application of factor analysis to economics, sociology, physiology, etc. is just beginning to get under way.

Quantitative psychology, which could be reasonably adequately covered by a six weeks' course in 1929, has moved a long way in the directions indicated by Thurstone (1937) in his Dartmouth address as retiring first

president of the Psychometric Society, "Psychology as a Quantitative Rational Science."

Presenting the topic, "Psychometrics--whence and whither," to this group is carrying coals to Newcastle, or maybe it is even gilding the lily, as you prefer. Much has been of necessity omitted in this brief presentation of a 40-year history. I have presented a few of the highlights, as I see them, and would be especially interested in your reactions.

References

Arbuckle, J. L. (1970) Multivariate analysis with stationary error: An application to auditory evoked potentials. Doctoral dissertation, Princeton University, and ETS Research Bulletin 70-36. Princeton, N. J.: Educational Testing Service.

Audley, R. J., & Jonckheere, A. R. (1956) Stochastic processes for learning. British Journal of Statistical Psychology, 9, 87-94.

Bass, F. M., King, C. W., & Pessemier, E. A. (Eds.) (1968) Application of science in marketing management, Purdue Symposium, 1968 Proceedings. New York: Wiley.

Bennett, G. E., Seashore, H. G., & Wesman, A. G. (1959) Manual for Differential Aptitude Tests. New York: Psychological Corporation.

Bennett, G. E., Seashore, H. G., & Wesman, A. G. (1966) Manual for Differential Aptitude Tests. New York: Psychological Corporation.

Best, P. J. (1966) An experimental investigation of a mathematical learning model. ONR Technical Report and doctoral dissertation, Princeton University, and ETS Research Bulletin 66-15. Princeton, N. J.: Educational Testing Service.

Binet, A. (1909) Les idees modernes sur les enfants. Paris: Ernest Flammarion. 344 pp.

Bock, D. R., & Jones, L. V. (1968) Measurement and prediction of judgment and choice. San Francisco: Holden Day. 370 pp.

Buckingham, B. R. (1921) Intelligence and its measurement: A Symposium. Journal of Educational Psychology, 12, 271-275.

Bush, R. R., & Estes, W. K. (1959) Studies in mathematical learning theory. Stanford, Calif.: Stanford University Press. viii + 432 pp.

Bush, R. R., & Mosteller, F. (1955) Stochastic models for learning. New York: Wiley. xvi + 365 pp.

Carroll, J. B. (1971) Measurement properties of subjective magnitude estimates of word frequency. Journal of Verbal Learning and Verbal Behavior, 10, 722-729.

Carroll, J. D., & Chang, J. (1970) Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika, 35, 283-319.

Carterette, E. C., & Friedman, M. P. (Eds.) (1973) Handbook of perception. New York: Academic Press.

Coombs, C. H. (1964) Theory of data. New York: Wiley. xviii + 585 pp.

Cronbach, L. J., & Gleser, G. C. (1965) Psychological tests and personnel decisions. (2nd ed.) Urbana: University of Illinois Press.

Dearborn, W. F. (1921) Intelligence and its measurement: A symposium. Journal of Educational Psychology, 12, 210-212.

Ebbinghaus, H. (1885) Über das Gedächtniss. Leipzig: Duncker & Humblat. ix + 169 pp.

Estes, W. K. (1956) The problem of inference from curves based on group data. Psychological Bulletin, 53, 134-140.

French, J. W. (1951) The description of aptitude and achievement tests in terms of rotated factors. Psychometric Monograph No. 5. Richmond, Va.: William Byrd Press.

Friendly, M. L. (1972) Proximity analysis and the structure of organization in free recall. ONR Technical Report and doctoral dissertation, Princeton University, and ETS Research Bulletin 72-3. Princeton, N. J.: Educational Testing Service.

Gagné, R. M. (1967) Learning and individual differences. Columbus, Ohio: Charles E. Merrill Books. xv + 265 pp.

Green, P. E., & Carmone, F. J. (1970) Multidimensional scaling and related techniques in marketing analysis. Boston, Mass.: Allyn and Bacon. xv + 203 pp.

Green, P. E., & Rao, V. R. (1972) Applied multidimensional scaling: A comparison of approaches and algorithms. New York: Holt, Rinehart, and Winston. xviii + 292 pp.

Gulliksen, H. (1934) A rational equation of the learning curve based on Thorndike's Law of Effect. Journal of General Psychology, 11, 395-434.

Gulliksen, H. (1942) An analysis of learning data which distinguishes between initial preference and learning ability. Psychometrika, 7, 171-194.

Gulliksen, H. (1950) Intrinsic validity. American Psychologist, 5, 511-517.

Gulliksen, H., & Gulliksen, D. P. (1971) Attitudes of different groups toward work, aims, goals, and activities. ONR Technical Report, Princeton University. Princeton, N. J.: Educational Testing Service.

Guttman, L. (1971) Measurement as structural theory. Psychometrika, 36, 329-347.

Harman, H. H. (1967) Modern factor analysis. (2nd ed.) Chicago: University of Chicago Press. 474 pp.

Harris, C. W. (1962) Some Rao-Guttman relationships. Psychometrika, 27, 247-263.

Harris, C. W. (Ed.) (1963) Problems in measuring change. Madison: University of Wisconsin Press.

Harris, G. G. (May 1972) Bibliography--multidimensional scaling (1960-1971). Bell Laboratories No. 205.

Helm, C. E., & Tucker, L. R (1962) Individual differences in the structure of color perception. American Journal of Psychology, 75, 437-444.

Horst, P. (1961) Generalized canonical correlations and their applications to experimental data. Journal of Clinical Psychology, Monograph Supplement No. 14, 331-347.

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, 417-441. 489-520.

Humphreys, L. G. (1943) Measures of strength of conditioned eyelid response. Journal of General Psychology, 29, 101-111.

Jöreskog, K. G. (1970) A general method for analysis of covariance structures. Biometrika, 57, 239-257.

Kaiser, H. P. (1970) A second generation little jiffy. Psychometrika, 35, 401-415.

Kelley, H. P. (1964) Memory abilities: A factor analysis. Psychometric Monograph No. 11. Richmond, Va.: William Byrd Press. 51 pp.

Kelley, T. L. (1939) Mental factors of no importance. Journal of Educational Psychology, 30, 139-142.

Kristof, W. (1972) An extension of the reliability concept to vector variables. Research Bulletin 72-7. Princeton, N. J.: Educational Testing Service.

Kuder, G. F., & Richardson, M. W. (1937) The theory of the estimation of test reliability. Psychometrika, 2, 151-160.

Leacock, S. B. (1911) Gertrude the governess: Or simple seventeen. In S. B. Leacock Nonsense novels. New York: John Lane. Pp. 7-230.

Lord, F. M. (1971) A theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 31, 805-813.

Lord, F. M., & Novick, M. R. (1968) Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley. xvii + 568 pp.

Luce, R. D., Bush, R. R., & Galanter, E. (1963a) Handbook of mathematical psychology, Vol. I and Vol. II. New York: Wiley.

Luce, R. D., Bush, R. R., & Galanter, E. (1963b) Readings in mathematical psychology, Vol. I. New York: Wiley.

Luce, R. D., Bush, R. R., & Galanter, E. (1965a) Handbook of mathematical psychology, Vol. III. New York: Wiley.

Luce, R. D., Bush, R. R., & Galanter, E. (1965b) Readings in mathematical psychology, Vol. II. New York: Wiley.

Luce, R. D., & Tukey, J. W. (1964) Simultaneous conjoint measurement: A new type of fundamental measurement. Journal of Mathematical Psychology, 1, 1-27.

McDonald, R. P. (1967) Nonlinear factor analysis. Psychometric Monograph No. 15. Richmond, Va.: William Byrd Press. (Also doctoral dissertation, University of New England, Australia, 1963.)

Meikle, T. H., Jr., & Sechzer, J. A. (1960) Interocular transfer of brightness discrimination in "split brain" cats. Science, 132, 734-735.

Meikle, T. H., Jr., Sechzer, J. A., & Stellar, E. (1962) Interhemispheric transfer of tactile conditioned responses in corpus callosum sectioned cats. Journal of Neurophysiology, 25, 530-543.

Merrell, M. (1931) The relationship of individual growth to average growth. Human Biology, 3, 37-70.

Messick, S. (1956) The perception of social attitudes. Journal of Abnormal and Social Psychology, 52, 57-66.

Müller, G. E., & Schumann, R. (1894) Experimentelle Beitrage zur Untersuchung des Gedächtnisses. Zeitschrift für angewandte Psychologie, 6, 81-190, 257-339.

Peterson, J. (1926) Early conceptions and tests of intelligence. Yonkers-on-Hudson: World Book. xiv + 320 pp.

Plotkin, L. (1972) Coal handling, steamfitting, psychology, and law. American Psychologist, 27, 202-204.

Pyle, W. H. (1921) The psychology of learning. Baltimore: Warwick & York.

Ramsay, J. O. (1970) A family of gradient methods for optimization. The Computer Journal, 13, 413-417.

Rose, R. M., Beach, L. R., & Peterson, C. R. (1971) Failures to find probability matching. ONR Technical Report, Number 71-1-19 (October 31, 1971).

Russell, I. S., & Kleinman, D. (1970) Task difficulty and lateralization of learning in the functionally split-brain rat. Physiology and Behavior, 5, 469-478.

Schiffman, H., & Falkenberg, P. (1968) The organization of stimuli and sensory neurons. Physiology and Behavior, 3, 197-201.

Shepard, R. N., Romney, A. K., & Nerlove, S. B. (Eds.) (June 1972) Multidimensional scaling: Theory and applications in the behavioral sciences. Vols. I and II. New York: Seminar Press.

Sidman, M. (1952) A note on functional relations obtained from group data. Psychological Bulletin, 49, 263-269.

Simrall, D. (1947) Intelligence and the ability to learn. Journal of Psychology, 23, 27-43.

Sperry, R. W. (1961) Cerebral organization and behavior. Science, 133, 1749-1757.

Sperry, R. W. (1964) The great cerebral commissure. Scientific American, 210, 42-52.

Stake, R. E. (1961) Learning parameters, aptitudes and achievements. Psychometric Monograph No. 9. Richmond, Va.: William Byrd Press. 70 pp.

Stuit, D. B. (Ed.) (1947) Personnel research and test development in the Bureau of Naval Personnel. Princeton, N. J.: Princeton University Press. xxiv + 513 pp.

Thorndike, E. L. (1918) The nature, purposes, and general methods of measurements of educational products. In The Seventeenth Yearbook of the National Society for the Study of Education, Part II, the Measurement of Educational Products. Bloomington, Illinois: The Public School Publishing Co. Pp. 16-24.

Thurber, J. (1937) Sex ex machina. In J. Thurber Let your mind alone. (3rd ed.) New York: Harper and Brothers.

Thurstone, L. L. (1930a) The learning function. Journal of General Psychology, 3, 469-493.

Thurstone, L. L. (1930b) The relation between learning time and length of task. Psychological Review, 37, 44-53.

Thurstone, L. L. (1931a) Multiple factor analysis. Psychological Review, 38, 406-427.

Thurstone, L. L. (1931b) A multiple factor study of vocational interests. Personnel Journal, 10, 198-205.

Thurstone, L. L. (1931c) The reliability and validity of tests. Ann Arbor: Edwards Brothers. 113 pp.

Thurstone, L. L. (1933) The theory of multiple factors. Chicago: University of Chicago Press. 65 pp.

Thurstone, L. L. (1935) The vectors of mind. Chicago: University of Chicago Press. 266 pp.

Thurstone, L. L. (1937) Psychology as a quantitative rational science. Science, 85, 228-232.

Thurstone, L. L. (1947) Multiple factor analysis. Chicago: University of Chicago Press. 535 pp.

Thurstone, L. L. (1959) The measurement of values. Chicago: University of Chicago Press. 322 pp.

Torgerson, W. S. (1951) A theoretical and empirical investigation of multidimensional scaling. Doctoral dissertation, Princeton University, and ETS Research Bulletin 51-14. Princeton, N. J.: Educational Testing Service. (Also published in Psychometrika, 1952, 17, 401-419.)

Torgerson, W. S. (1958) Theory and methods of scaling. New York: Wiley.

Tryon, R. C. (1955) Identification of social areas by cluster analysis (A general method with an application to the San Francisco Bay Area). University of California Publications in Psychology, 8(1), 1-100. Berkeley: University of California Press.

Tucker, L. R (1956) Factor analysis of double centered score matrices. Research Memorandum 56-3. Princeton, N. J.: Educational Testing Service.

Tucker, L. R (1958) An inter-battery method of factor analysis. Psychometrika, 23, 111-136.

Tucker, L. R (1966a) Some mathematical notes on three-mode factor analysis. Psychometrika, 31, 279-311.

Tucker, L. R (1966b) Learning theory and multivariate experiment: Illustration by determination of generalized learning curves. In R. B. Cattell (Ed.), Handbook of multivariate experimental psychology. Chicago: Rand McNally. Pp. 476-501.

Tucker, L. R (1972) Relations between multidimensional scaling and three-mode factor analysis. Psychometrika, 37, 3-27.

Tucker, L. R, & Messick, S. (1963) An individual differences model for multidimensional scaling. Psychometrika, 28, 333-367.

Wainer, H. (1968) A principal components analysis of models and men. ONR Technical Report and doctoral dissertation, Princeton University, and ETS Research Bulletin 68-29. Princeton, N. J.: Educational Testing Service.

Weitzman, R. A. (1963) A factor analytic method for investigating differences between groups of individual learning curves. Psychometrika, 28, 69-80.

Wissler, C. (1901) The correlation of mental and physical tests. Psychological Review, Monograph Supplement 3, No. 6.

Woodrow, H. A. (1946) The ability to learn. Psychological Review, 53, 147-158.

Young, G., & Householder, A. (1938) A discussion of a set of points in terms of their mutual distances. Psychometrika, 3, 19-22.